Benchmarks as Microscopes: A Call for Model Metrology

Michael Saxon[*1] **Naomi Saphra**[*2] **Ari Holtzman**[*3] **Peter West**[*4] **William Yang Wang**[1]
[1]UC Santa Barbara [2]Harvard University [3] University of Chicago [4]University of Washington

Abstract

Modern language models (LMs) have made AI system capability assessment harder than ever before. Static benchmarks inevitably saturate without providing confidence in the deployment tolerances of these systems, but developers nonetheless claim generalized traits such as reasoning or open-domain language understanding based on these flawed metrics. The science and practice of LMs requires a new approach to benchmarking which measures specific capabilities with dynamic assessments. In other words, we need a new discipline of *model metrology*—one which focuses on how to generate benchmarks that predict performance under deployment.

In this position paper, we describe these desiderata in detail, explain why they are necessary, and outline how building a community of model metrology practicioners—primarily focused on building tools and understanding of how to measure AI system capabilities—is the best way to meet these needs, and add clarity to the AI discussion.

1 Introduction

Just how good are our current language models? It's hard to say. Even the latest benchmarks are not scalable, relevant, or durable enough to predict performance in real-world settings.

Although engineering (Saka et al., 2024; Goyal et al., 2024), research (Bommasani et al., 2021; Bai et al., 2022), and policy (Cihon, 2019; NIST, 2023) decisions are being made based on claims around AI—particularly LM (Tolan et al., 2021; NIST, 2022)—capabilities, there is fundamental contention about the nature (Li et al., 2023; Morris et al., 2023) and extent (Jumelet & Hupkes, 2018; Bender & Koller, 2020; Park et al., 2022) of these capabilities. Popular assessments of LM breakthroughs rely on either **aggregate performance on many narrow benchmarks** (Achiam et al., 2023) or **one-off manual analyses** (Bubeck et al., 2023); these assessments then form the basis of our conversation around scaling (Hoffmann et al., 2022), generalization, risk (Falco et al., 2021), and deployability. Popular static benchmarks inevitably *saturate* (Beyer et al., 2020; Kiela et al., 2021; Ott et al., 2022) as each consecutive generation of models over-optimizes for performance on its evaluation sets—a process exacerbated by those datasets contaminating future training data—without resolving these fundamental impasses over the nature of LMs or usefully informing their deployment in critical applications. *We need benchmark practices that yield meaningful scientific observations.*

The emergence of a new discipline: from homemade microscopes to optical metrology

In 1609, Galileo Galilei built one of the first optical telescopes. By turning it around, Galileo found that he could also observe very small objects close up—and so microscopy was born (Singer, 1914). For centuries, these tools were built by the same scientists using them to make fundamental discoveries (La Berge, 1999).

Eventually, the expertise required to to design increasingly complex and precise tools outstripped scientists' skill at the craft. In the 20th century, massive radio telescopes and orbital space telescopes were built by large teams of specialists (Leverington, 2012), while microscopes became commodity products from dedicated firms (Davidson & Abramowitz, 2002). For both tools, the science and engineering practices of measurement have coalesced into specialized disciplines unto themselves.

[*] Equal contribution; corresponding author: saxon@ucsb.edu

At present, our LM evaluation practices resemble the state of astronomy and microbiology in the early 17$^{\text{th}}$ century—the same community analyzing the object of study (models) is also building the tools for that analysis (benchmarks). We believe that within the study of LMs, the practice developing analytical tools must advance similarly to the practice of microscope and telescope building in the 20$^{\text{th}}$ century. An independent discipline focused on LM evaluation tools must emerge to realize this transition. We call for formalizing this new discipline, *model metrology*, devoted to building evaluations which are **dynamically generated**, **constrained** to a specific task or domain, **ecologically valid** in their reflection of deployment conditions, and **plug-and-play** to permit rapid evaluation of new models, based on meaningful rather than arbitrary observations. In short:

- There exists a gap between what is testable and what LMs are actually used for, a gap which current benchmarks fail to close.
- To close this gap, evaluation must be dynamic, constrained, ecologically valid, and plug-and-play.
- A new discipline, *model metrology*, is needed to advance the theory and practice of building benchmarks that meaningfully reflect deployment use-cases.

2 Problems with current benchmarks for LMs

Benchmarks have long driven AI advances. When research communities believe that "solving" a benchmark represents core progress toward generalized intelligence, it is natural that interest and investment towards that end follow. Raji et al. (2021) document how the *common task framework*—public contests between systems assessed through quantitative evaluation over common train and test sets (Donoho, 2017)—enabled advancements in concrete and tightly-scoped problems such as automatic speech recognition and machine translation, but **has since been inappropriately extended to claim generalized capabilities** of vision (Russakovsky et al., 2015) and language (Wang et al., 2019) models.

The LLM age has seen a shift from training and testing directly on benchmark datasets to testing models in a zero-shot setting. Consequently, researchers often assume that *because* LMs aren't explicitly trained on task-specific train sets (Ge et al., 2023; Bai et al., 2024), performance on these benchmarks is *stronger evidence* of general capability than for fine-tuned models (Piantadosi & Hill, 2022; Mitchell & Krakauer, 2023). In general, perspectives on these questions among NLP and AI researchers are remarkably divergent (Michael et al., 2023). This lack of clarity is bad enough for practitioner evaluation of competing systems, but escalating claims of general intelligence have also drawn public attention (Neri & Cozman, 2020) in debates over AI risk (Ambartsoumean & Yampolskiy, 2023).

Such attempts to assess general capabilities come from a legitimate need: evaluation is important for guiding advances and comparing models (Phillips et al., 2000). Because LLMs are used as generalized *everything systems*, a desire to characterize their general capabilities naturally follows (Morris et al., 2023). **Current evaluation methods, in attempting to assess general capabilities, fail to meaningfully capture either general competency or the specificity required for functional target applications**.

2.1 Generalized capabilities are hard to define and contentious

Narrow LM capability benchmarks are often derived either from *tests for humans*—e.g., GSM8K (Cobbe et al., 2021) and MMLU (Hendrycks et al., 2020)—or from *existing constrained benchmarks* for specific engineering problems in natural language processing, such as question answering (Ho et al., 2023) or natural language inference (Bowman et al., 2015). Within a narrow scope, models which excel on these benchmarks likely generalize to similar test sets on the same task, as evidenced by off-trend behavior for this correspondence being considered indicative of test-set contamination (Paster, 2023; Jain et al., 2024, fig. 5). However, claims regarding the real-world reliability of these systems are often poorly supported (Liao et al., 2021). Furthermore, these benchmarks cannot characterize broader capabilities (Srivastava et al., 2023), even when aggregated (Liang et al., 2023).

Critics of generalized capabilities benchmarks note that they lack *construct validity*—in short, strong demonstration that any evaluation represents a capability (O'Leary-Kelly & Vokurka, 1998) to support research claims around it. This issue around generalized capabilities was already present for fine-tuned model evaluation (Raji et al., 2021), and has since worsened, as LM developers and their allies claim generalized intelligence (Bubeck et al., 2023) using collections of scores across a huge set of limited benchmarks (Fei et al., 2022; Achiam et al., 2023). Similar claims are made for specific abstract capabilities; benchmark performance is attributed to faculties like abstract reasoning (Yasunaga et al., 2021), language understanding (Moore, 2022), or common sense knowledge (Zhao et al., 2023b).

Yet these claims are contested. Rebuttals include axiomatic arguments against the possibility of acquiring these capabilities from language modeling objectives (Bender & Koller, 2020). Results on poor generalization across times (Lazaridou et al., 2021), tasks (Yang et al., 2022) and heuristics (Singhal et al., 2022) have also been provided as empirical counterarguments to generalized capabilities in current systems. Abstract notions like reasoning are fundamentally slippery (Manning, 2022) and humans, who cannot even affectively evaluate intelligence in our fellow animals (De Waal, 2016), may similarly fail for complex AI systems.

On the wrasse fish & the pitfalls of generalized capabilities

As the apocryphal Einstein quote goes, "if you judge a fish by its ability to climb a tree, it will live its whole life believing that it is stupid" (O'Toole, 2017).

Ironically, there is at least one example of fish intelligence outpacing primates, namely the economic puzzles solved in labs by the *cleaner wrasse*, a symbiotic species that lives in coral reefs, feeding on the parasites of larger fish. The fish are given meal options which should be eaten in a particular order due to variable reliability, and they find the optimal solution faster than capuchins, chimpanzees, orangutans, and even one researcher's four-year-old daughter (Salwiczek et al., 2012).

The wrasse evolved to preferentially treat regular customers over reef visitors, tracking clientele across thousands of daily parasite cleanings (Gibson & Barnes, 2000). An artificial ***wrassebot*** cleaner would be ready to deploy only when it exhibits game theoretically-optimal strategies and "machiavellian" (Bshary, 2011) manipulations of its clientele; there's no need to solve grade school math problems or translate text.

2.2 Desire for generality in benchmarks hurts validity & utility

Lurking under the surface of many critiques of benchmarking general capabilities is this question: is the desired abstract capability **necessary** or just **sufficient** for solving the benchmark task (Potts, 2020)? For any benchmark to capture an abstract capability—even setting aside the aforementioned problems of construct validity—the capability *must be necessary* to solve the evaluation. It turns out that this is a massive challenge. After all, an answer key can "solve" an exam with 100% accuracy, as can a well-studied human, and it is clear that the answer key does not possess intelligence in the way that a human does.

However, **for concrete applications, all a useful benchmark must do is demonstrate that a system**—using whatever combination of capabilities—**has the sufficient capabilities to perform the task**, regardless of abstract mechanisms. This was the original motivation for shared tasks (Raji et al., 2021).

Unfortunately, the contrived nature of attempts at generality also costs *ecological validity* (De Vries et al., 2020), meaning behaviors of models on these examples are likely unrepresentative of in-the-wild behavior. Recent efforts to overcome this problem such as HELM (Liang et al., 2023) provide a large collection of scores by harvesting existing benchmarks and bucketing them by specific scenario (e.g., news domain tasks). This contextualization does enable more useful comparisons (Liao & Xiao, 2023), but each benchmark represents a tiny view of a broader "task universe" (Liao et al., 2021). Ecological validity is even a problem in benchmarking "AGI-level" capabilities (Morris et al., 2023), though discussing the credibility of those notions is not central here.

Most model consumers (i.e., developers of applications that rely on predictable LM behavior) need a laser-focused evaluation of consistent, constrained capabilities for their domain. However, developers such as OpenAI (Achiam et al., 2023), Anthropic (Anthropic, 2023), and Google (Team et al., 2023; 2024) instead continue to focus the same MMLU, GSM8K, and HumanEval test suites, rather than branch out to application-specific tests, or take a more holistic approach as in HELM.

Seeing as these groups provide API access to these models as a paid service, **why aren't they benchmarking customer-relevant capabilities?** One reason might be that they still cling to the idea that general intelligence is quantifiable by these benchmarks. Perhaps the deeper issue is that building bespoke evaluations is hard, and the domains are innumerable—will a collection of constrained tests ever be large enough to placate critics?

2.3 Static benchmarks inevitably die by saturation, so why keep using them?

Goodhart's law states roughly that "when a metric becomes a target, it loses value as a metric" (Goodhart, 1984). AI research demonstrates this law quite well; it's a near truism that static benchmarks saturate so fast that they are effectively useless within a few years.

In an environment where models are being trained at scale on incomprehensibly large proprietary datasets (Elazar et al., 2023), the impact of legitimacy-challenging effects such as memorization (Alzahrani et al., 2024) via leakage of similar examples in the long tail of training data (Peng et al., 2023) are still questioned in debate (Kandpal et al., 2023). It is well-documented that language models learn heuristic solutions (Poliak et al., 2018; McCoy et al., 2019) in the supervised setting (Wang et al., 2021; Saxon et al., 2023). It is well worth considering that benchmark-specific heuristics and memorization play a significant role in model performance on static benchmarks. In light of this, **what useful insight is gained when a model saturates a benchmark?** In the case of GSM8K, the only new approach required was a combination of prompting tricks and decoding techniques that brought GPT-4 to near-100% accuracy (Zhou et al., 2024a). Why should we expect the case of MMLU to be any different when it is inevitably "solved?"

3 Qualities of useful, concrete benchmarks

In light of the above issues, is it a surprise that many view extant LM evaluations as useless for deployment decisions? We believe the solution is to **abandon the futile pursuit of measuring generality**. **Why study the *wrasse fish*'s intelligence outside of the reef?**

We need a benchmarking culture and practice that *empowers consumers to specify their evaluation constraints* and *generate their own benchmarks directly*. Model producers should track progress by these scenario-driven evaluations. **We believe good LM evaluations are:**

Constrained. Benchmarks that focus on characterizing model behavior with respect to a concrete task, where customers or domain experts can describe the boundaries and limits a good system should stay within are *constrained*. Often, concrete problems can be defined in terms of direct "*don't do x*"-type constraints. Within these concrete domains, issues of scenario innumerability and *ecological validity* become much more tractable—domain experts have a better understanding of real-world inputs than AI researchers.

Dynamic. One reason is that static benchmarks of that form are easily memorized, warranting the production of *dynamically generated benchmarks.* Models can't directly memorize the answers, lending more credence to the generality of performant results. In constrained settings generators can be more readily characterized by task-specific constraints (e.g., *don't give a fake discount*) than in general settings (e.g., *don't disobey rules*).

Plug-and-play. Benchmarks and generating processes that can be run easily using customer specific models or constraints are *plug-and-play*. Benchmarks which are inaccessible or infeasible to use for a vast majority of the community are not useful in creating common ground. Expensive, one-off benchmarks do have a place in well-funded organizations, but they are unlikely to foster the shared reasoning points and open discourse that academic research aims for. Plug-and-play setups will entice more diverse contributors of constraints.

These desiderata are mutually reinforcing and enabled by *abandoning generality*.

4 Model metrology in an ideal world

> ***The Air Canada chatbot incident***
>
> A Canadian court found Air Canada liable for paying a customer a nonexistent, off-policy bereavement discount promised by an LM-based customer service agent (Melnick, 2024). This agent, presumably built with a GPT base model, failed to adhere to policy—thereby failing as a customer service agent—despite GPT-3.5's near state-of-the-art performance on a massive suite of generalized benchmarks.

Having made the case for why general assessment of abstract capabilities should be abandoned, and what good concrete benchmarks of constrained capabilities should look like, we must ask, **why aren't such benchmarks currently being built?** Three potential reasons are:

1. Lack of communication between model evaluation builders and model users
2. Misalignment of interest/incentives for researchers and needs of users
3. Fundamental difficulties in building benchmarks that meet our desiderata

In this section, we lay out the fundamental case for why building a dedicated benchmarking discipline—of *model metrology*—can resolve these discrepancies, producing benchmarks that meet our desiderata and improve the state of LM production, use, and analysis.

4.1 Closing the gap between researchers, developers, and users is crucial

We strongly believe that involvement of these stakeholders in LM benchmarking culture must grow alongside the development of the benchmarking discipline. Model metrology has to be a sociological development, in addition to a scientific one: incentives must exist for both metrologists and benchmark users to engage in productive scientific discourse.

Even if ideal tooling and methods existed already to expand user constraints into quality dynamic benchmarks (outlook and roadblocks to this discussed in §4.2), getting active and prospective LM-based application developers (the "users" of these models) to provide specifications that meaningfully capture their use case may still be a challenge.

For example, consider the aforementioned *Air Canada chatbot incident*. This incident could be described as the underlying LM lacking several different abstract capabilities. Perhaps the agent failed at *rule-following*, and the constraint of what the rules for bereavement discounts are was specified somewhere in the system prompt context. Perhaps the agent failed in *commonsense reasoning*; this specific policy wasn't explicitly posed, but the list of all allowable discounts was. Perhaps the agent failed at some third, different capability. Regardless of the source of the error, we find it implausible that this failure case could have been predicted through generalized benchmark results.

However, a developer of customer service agents specifically could have tested for this. Within the customer service agent domain, *common sense entails not making promises out of policy*. A developer could manually write test cases for every line of the policy, probing the agent for examples where it would fail. Model developers aren't thinking what abstract capability failures mean in diverse domains, *but domain experts who will use the models know what their needs are*. To truly know how good a model is in the real world, we need as many real world scenarios to test on as possible—this way we may achieve ecological validity.

A dedicated model metrology community will both be an audience for and consultants to model consumers who know what boundary conditions they need to test. Metrologists then can combine these tests into better holistic evaluations for use by model makers.

4.2 Building useful & deployable dynamic benchmarks for complex problems

What techniques might be employed in building evaluations that meet our desiderata? Why is a community of practice necessary to enable their development? In this section we these questions, grounded with examples inspired by the *Air Canada incident*.

Suppose a customer service agent developer had a concrete list of constraints, grounded in "*don't do x*" language, including not promising off-policy transactions. This is a great example of a **constrained** domain. How do we dynamically evaluate an LM for this task?

One technique could be to leverage an *adversarial LM* as a source of variation, generating many test cases attacking the task-domain constraints. For example, for the airline assistant role, an LM could be prompted to generate role-play scenarios of various customers asking a chatbot to be a child pranking the system, a client who struggles with technology, a jailbreaker looking for a big discount, or a panicked and angry stranded traveler.

Model outputs conditioned on these adversarial test inputs could be judged deterministically against policy constraints, and scenarios such as the lying discount behavior may be detected. The human evaluator doesn't need to manually enumerate all edge cases, so this benchmark isn't static. **Dynamic benchmarks are best produced in expert-constrained settings**.

There are reasons to believe this is possible. LMs have been used as adversaries to other LMs for stress testing and assessment (Chan et al., 2024). There is mounting evidence that LMs can generate exemplars that they can't "understand" via reversal (Berglund et al., 2023; West et al., 2023). Given well-scoped constraints, the outputs can be evaluated deterministically, eg via variation between minimal pairs (Ribeiro et al., 2020), or by individually assessed binary fulfillment of a set of requirements (Hu et al., 2023), rather than using arbitrary and opaque LM-judgements of dubious reliability (Oh et al., 2024).

Though we have hypothesized a pipeline to produce a dynamic benchmark for one constrained setting, we are not claiming this is the best way to produce a strong benchmark-generating process for all settings. Concerted research is necessary to develop best practices. Developing, formalizing, and sharing insights from disparate benchmark efforts this practice is one of the most important competencies model metrologists will have.

4.2.1 Static benchmarks as targets to evaluate benchmark generators

While static benchmarks are essentially stale on release, they still may have value as *prototype outputs of a benchmark-generating process*. In a setting where expert expectations and system constraints can clearly be defined, producing exemplars or simulating problematic interactions can be achieved through direct human effort. For current static benchmarks, small test sets (of say, 50-100 samples) can estimate the performance of a model on large test sets covering multiple tasks (Polo et al., 2024). Rather than using a static benchmark as-is to evaluate models, we can repurpose prototypical static benchmarks as **output targets** for the generating processes. This can enable *meta-evaluation of benchmark generation techniques.*

4.2.2 Meaningful saturation points to capture engineering tolerances

When we build tools, we often consider their *tolerance*, which for a single dimension output can be described as *the maximum variation from the target that the tool may randomly exhibit* (Mansoor, 1963). In LMs things cannot be so easily defined, as we do not have the luxury of a single dimension or even a fixed space within which to measure such variance. And yet this definition captures something that has stopped LMs from being as widely used as such a qualitatively different tool might otherwise have been: when applying an LM to a new task in a new domain, we have no widely-applicable way of predicting how much and what kind of variance to expect from an LM, *even after its results have been produced*.

This motivates the need to measure LMs where we believe they will go wrong: in constrained domains with specific goals. We can then look whether a bundle of domains can help us predict the tolerance in the next domain, but general measures will never give us this kind of precision, because they are purposefully smooth where real-world domains are disjoint and idiosyncratic. Another ideal quality for real-world benchmarks to have is a **meaningful saturation point**. Is it possible to achieve this through targeted benchmarks?

Rather than high agreement around one static correct answer, we instead focus on a list of negatives that the model needs to avoid producing. We believe this should be the goal. When an ideal benchmark saturates, all models at this accuracy should be considered **good deployment candidates**. Ironically, this benchmarking paradigm brings us back to the stable ground of the original common task framework (Liberman, 2010; Donoho, 2017).

4.3 The role of metrologists

Here we lay out the case for building model metrology *as a discipline*, based on the practices and understanding that are best developed through an organized community. Model metrologists use the theory and practice they develop to produce evaluation solutions for model consumers and improved benchmarks for model makers.

4.3.1 Establish shared knowledge & techniques

Core questions that could be very useful, such as whether LMs can provide meaningful scores to a text passage in terms of reference-similarity (Kocmi & Federmann, 2023) or correctness (Wang et al., 2023a; Mizumoto & Eguchi, 2023), have no consensus answer (Chiang & Lee, 2023; Oh et al., 2024). Arriving at a consensus on the efficacy of evaluation techniques would allow metrologists use them confidently rather than rely on blind faith.

Observation canonization pipeline The chaotic and fast-paced nature of LM experimentation presents a challenge for making use of disparate observations: *which observations are meaningful presentations of underlying phenomena, and which are just random one-off behaviors*? Observations are being generated by by the broader AI research community so fast that it's a challenge to keep up. One critical benefit of a dedicated discipline would be to prevent *repeat work* in surfacing relevant findings, by replicating, analyzing, and eventually "canonizing" empirical findings that are important knowledge for metrology practitioners across all domains. For example, we only have fleeting evidence suggesting connection between benchmarks on various domains (Fergusson et al., 2023). There are constantly competing hypotheses and evidence that e.g., subtasks of MMLU are usefully correlated (Paster, 2023; Jain et al., 2024), while others claim they aren't.

Shared framings of abstract capabilities in concrete settings While assessment of abstract capabilities like "reasoning," "understanding," or "rule-following" are especially slippery and innumerable in the open domain, they can be used to frame desired behavior and edge cases to avoid in constrained settings. It may be advantageous to think of these as *horizontals* with greater overlap in methodology. Through comparison of results of deployment of similar methods between disparate settings, metrologists will make useful scientific contributions that can guide further tooling development.

Benchmark-building tools An example of this tooling development might be finding prompting techniques that work well for testing the boundaries of rule-following in one setting (e.g., customer service) would generalize better to other settings (e.g., planning navigation), than do prompting techniques intended for reasoning assessment.

Eliciting agents to interact with systems will, in many cases be the best way to probe the constraints. Understanding optimal prompting techniques to elicit personae (Cheng et al., 2023) could be pivotal for the development of stress tests for interactive chatbot rule-following, while perhaps for commonsense reasoning in vision and language (Bitton-Guetta et al., 2023), retrieval of good contrastive images from the internet might be best.

4.3.2 Act as a go-between connecting model builders to application users

As covered above, most base model releases prioritize reporting scores on narrow, static benchmarks in a futile attempt to demonstrate the base model's generalized capabilities.

Dynamic but non-concrete benchmarks that have been proposed, e.g. Dynabench (Kiela et al., 2021), while free from the saturation weaknesses of static benchmarks, still attempts for the flawed generality strategy, using engineering problems like NLI, sentiment analysis, and hate speech detection as targets. As a result it has few benefits over the popular static benchmarks in modeling application-useful qualities, and is thus not used by model producers.

Similarly, *concrete, constrained, but static* evaluations do get proposed often—in methods papers, as one-off effective footnotes to demonstrate that the new method works. For example, Cho et al. (2023) produce a constrained test set for evaluating visual question answering accuracy, to assess the quality of VLMs for the generative image scoring task. But the lack of focus on the benchmark as a contribution leads it to be overlooked.

Model metrologists would develop an expertise on task- and domain-specific evaluation practices, both by producing evaluations in concert with domain experts, and by paying attention to where new evaluations are proposed in the modeling/intervention literature. In turn, when metrologists are engaged by model makers to assess the performance of new releases, they can bring these better evaluations to the table, ensuring more useful information is gained by model consumers, and more useful targets for advancement are given to model developers.

5 Building the model metrology discipline

Having established the motivation, purpose, and necessity of the dedicated discipline of model metrology, we now turn to a discussion of potential ways to build the field.

Multiple sociological changes to academic incentive structures and application developer engagement with model developers will be needed to realize the ideals of metrology.

5.1 Bringing proto-metrologists together

At present many conferences which deal with LMs and AI already have benchmarks and evaluation tracks. Work describing best practices and calling for benchmark development have appeared in these tracks at ICLR (Lu et al., 2024), *ACL (Maynez et al., 2023), CVPR (Xu et al., 2022), and NeurIPS (Zhang et al., 2023). The researchers engaged in these directions are effectively *model proto-metrologists*, as Galileo was in prefiguring optical metrology.

As a starting point, existing directions of work being contributed to these disparate venues can be treated as an initial canon that current and aspiring metrology researchers should be familiarized with. Non-archival workshops and dissemination venues could facilitate cross-engagement, and eventually archival venues on evaluation across all domains could be stood up to foster growth and give aspiring metrologists an intellectual home.

Additionally, future benchmarking and model performance analysis work appearing at these venues should explicitly call out its connection to the model metrology discipline, to drive visibility and interest.

5.2 Engaging with and distinguishing from related fields

Model metrology will be intimately connected to many subfields of machine learning, artificial intelligence, and natural language processing research, and will require engagement with them for its success. In particular, we anticipate work from subfields such as black-box model analysis (Belinkov et al., 2023), human-computer interaction (Liu et al., 2023), and mechanistic interpretability (Räuker et al., 2023), will be influential.

Model metrology will be distinct from these fields due to its focus on developing *observational tools* rather than *analytical tools*. For example, transformer circuits (Elhage et al., 2021) and induction heads (Olsson et al., 2022) are analytical tools of mechanistic interpretability that are used to study observations. Benchmarks of the type we call for are instead *observational tools*, which will yield phenomena that analytical work can investigate with analytical tools.

5.3 Soliciting novel constraints and edge cases to benchmark

Building the model metrology community requires deep engagement with domain experts and model users. Using hypothetical settings such as the Air Canada incident are a reasonable starting points for practitioners to experiment with constrained benchmark-generating systems, but ultimately **useful constraints can only be provided by domain experts** who know the boundaries of their task. Academic metrologists and metrology venues should actively solicit constraints for new task areas. These could be framed as *shared tasks* for task areas, or even as *concrete eval bounties.*

We think it's likely that industry and nonprofit "customers" of LM-based applications will be happy to contribute their constraint specs—after all they stand to benefit a lot if their needs are factored in by model training institutions.

It's likely that *industrial model metrology* will emerge as a profitable area for both startups and individuals as "metrology contractors." Experts in the measurement of models will be invaluable assets to firms building LM-modulo systems (Kambhampati et al., 2024).

6 Conclusions

Model metrology is a promising blue sky discipline. If successful, the agenda will enable more sensible and broadly agreeable ways to frame and operationalize many areas of interest within AI, including alignment, common sense, knowledge, and reasoning, by freeing them from the innumerability and construct validity issues endemic to treating them as generalized, open-domain capabilities. It will produce real-world applicable evaluation techniques that will help LM users make more informed decisions and enable model developers to track progress in a more fine-grained manner. We believe using grounded benchmarks to track progress will also improve public discourse around AI.

6.1 Model metrology & the aritficial general intelligence (AGI) discussion

Present benchmarking culture represents a clash between a practical need to compare models and ideological debates about AGI. The AGI-credulous decry every "goalpost move" from a newly-saturated benchmark to a different one in support of a "deep learning hitting a wall" narrative, despite a total lack of evidence that the capabilities of interest are necessary for saturation on those benchmarks. At the same time, the AGI-skeptical will make premature claims about the impossibility of LMs having a capability means they will never score high on some benchmark, the benchmark saturates, and the cycle continues.

We believe our vision for metrology is useful as it directly gets at a core reason most people actually care about the AGI: the promise of making drop-in replacements for humans in specific jobs. If this really is what we care about, **why not measure it directly**? This is the purpose of **constrained** and **ecologically valid** benchmarking.

A culture of model metrology will hopefully drive everyone to make **weaker statements** about intrinsic model capabilities grounded in quantifiable real-world capacities, and promote a healthier (and less panicked) public-facing discourse around AI.

6.2 The end game: model assessment *without* a model metrologist

In the best-case scenario, success of the model metrology agenda entails turning it into a mature engineering discipline, and making standardized benchmark-generating processes effectively off-the-shelf industrial goods, like how microscopes are today.

For most applications requiring a microscope, the exact desired instrument is already mass-produced. For truly niche applications (e.g., assessing a specific kind of semiconductor deformation) custom-built metrology solutions are still needed (Houghton et al., 2016).

One day, LM consumers should similarly be able to meet their measurement needs out-of-the-box without hiring a metrologist; the model metrology discipline will get us there.

7 Ethics statement (limitations)

As this work doesn't introduce new techniques or data, a traditional ethics statement is unwarranted. Instead we discuss our proposal's limitations, posed as rhetorical questions and comments.

You overlooked existing dynamic/constrained/construct valid benchmark examples! While automatically-generated benchmarks have already been proposed, they have been confined to mostly toy problems of limited interested to practicioners (with one notable counterexample being code generation). For these tasks, games like chess (Feng et al., 2023), reasoning and planning problems from block worlds (Valmeekam et al., 2022; Stechly et al., 2024), and deterministically evaluable natural domains such as programming (Allamanis et al., 2024), test examples can be generated deterministically. Unfortunately, apart from programming these are toy problems with limited insights in useful applications, and even for programming, static benchmarks (e.g., HumanEval (Chen et al., 2021)) reign supreme. Metrologist effort will enable more dynamic benchmarks reflecting real applications.

As for constrained and construct valid benchmark examples we didn't discuss, WildBench (Lin et al., 2024) is a static benchmark built atop exemplars collected from WildChat (Zhao et al., 2023a). This is one rare example of a benchmark that truly is representative of its target distribution (built from real user interactions with GPT!) However, it still is static.

> ***Reflections on trusting trust & self-verification of complex systems***
>
> Ken Thompson's Turing award acceptance speech, "Reflections on Trusting Trust" (Thompson, 1984), details how he hid a backdoor Trojan horse in early source versions of a C compiler. Because the compiler was bootstrapped, i.e., new versions of a compiler were compiled by the previous version, this backdoor was nearly impossible to detect or remove without being aware of its introduction.
>
> The backdoor was included even in versions of the compiler binary built from source code without the backdoor, as long as that source was compiled using a binary descended from Thompson's modified code. His discovery, shocking in 1984, seems almost mundane today: *If any part of a complex system is compromised, the entire system is compromised.* For modern automated metrics employing blackbox language models for their own evaluation, verification, and even training, we must view each element as potentially compromised by the flaws in proprietary models.

LMs evaluating LMs? How can a system measure capabilities we don't know it has? Relying on the target of analysis to self-verify has always been an issue. There's a chicken-or-egg problem where we have to ask: *how can we use a model to measure capabilities its own (or those of similar models)*? As we mentioned above, one solution is to expand out a set of more objectively evaluable characteristics, such as a set of yes/no questions (Hu et al., 2023) rather than subjective judgements by a model.

Preliminary evidence collected using human annotators suggests that even GPT-4, widely considered to be the most performant LM at submission date, has severe performance issues relative to humans in open-domain claim verification (Wang et al., 2023b).If we use prompt a model to generate sentences, then ask GPT-4 to evaluate the generated text, how can we trust that GPT-4's judgements capture anything meaningful? Subjective LM eval will lose the comparability of most benchmarks, as the particular model used for evaluation will affect the outcome (Zhou et al., 2024b).

Even dynamic benchmarks will go stale Living benchmarks based on silly/arbitrary rules can be gamed, learned by doing enough attempts on the eval. We have no visibility into the decision processes (Oh et al., 2024). This is part of why developing a living community is so important. New methods for producing benchmark generators will need to be refreshed. Breakthroughs may sometimes occur where generative techniques become obviated.

We leave all aforementioned issues to future work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Miltiadis Allamanis, Sheena Panthaplackel, and Pengcheng Yin. Unsupervised evaluation of code llms with round-trip correctness. *arXiv preprint arXiv:2402.08699*, 2024.

Norah Alzahrani, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yusef Almushaykeh, Faisal Mirza, Nouf Alotaibi, Nora Altwairesh, Areeb Alowisheq, et al. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. *arXiv preprint arXiv:2402.01781*, 2024.

Vemir Michael Ambartsoumean and Roman V Yampolskiy. Ai risk skepticism, a comprehensive survey. *arXiv preprint arXiv:2303.03885*, 2023.

Anthropic. The claude 3 model family: Opus, sonnet, haiku. Online technical report., 2023. URL `https://api.semanticscholar.org/CorpusID:268232499`.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. Benchmarking foundation models with language-model-as-an-examiner. *Advances in Neural Information Processing Systems*, 36, 2024.

Yonatan Belinkov, Sophie Hao, Jaap Jumelet, Najoung Kim, Arya McCarthy, and Hosein Mohebbi (eds.). *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, Singapore, December 2023. Association for Computational Linguistics. URL `https://aclanthology.org/2023.blackboxnlp-1.0`.

Emily M Bender and Alexander Koller. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 5185–5198, 2020.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: Llms trained on" a is b" fail to learn" b is a". *arXiv preprint arXiv:2309.12288*, 2023.

Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2616–2627, 2023.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, 2015.

Redouan Bshary. Machiavellian intelligence in fishes. *Fish cognition and behavior*, 2:240–257, 2011.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better LLM-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=FQepisCUWu.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. Compost: Characterizing and evaluating caricature in llm simulations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10853–10875, 2023.

Cheng-Han Chiang and Hung-Yi Lee. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15607–15631, 2023.

Jaemin Cho, Yushi Hu, Jason Michael Baldridge, Roopal Garg, Peter Anderson, Ranjay Krishna, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. In *The Twelfth International Conference on Learning Representations*, 2023.

Peter Cihon. Standards for ai governance: international standards to enable global coordination in ai research & development. *Future of Humanity Institute. University of Oxford*, pp. 340–342, 2019.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Michael W Davidson and Mortimer Abramowitz. Optical microscopy. *Encyclopedia of imaging science and technology*, 2(1106-1141):120, 2002.

Harm De Vries, Dzmitry Bahdanau, and Christopher Manning. Towards ecologically valid research on language user interfaces. *arXiv preprint arXiv:2007.14435*, 2020.

Frans De Waal. *Are we smart enough to know how smart animals are?* WW Norton & Company, 2016.

David Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26 (4):745–766, 2017.

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, et al. What's in my big data? *arXiv preprint arXiv:2310.20707*, 2023.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1:1, 2021.

Gregory Falco, Ben Shneiderman, Julia Badger, Ryan Carrier, Anton Dahbura, David Danks, Martin Eling, Alwyn Goodloe, Jerry Gupta, Christopher Hart, et al. Governing ai safety through independent audits. *Nature Machine Intelligence*, 3(7):566–571, 2021.

Nanyi Fei, Zhiwu Lu, Yizhao Gao, Guoxing Yang, Yuqi Huo, Jingyuan Wen, Haoyu Lu, Ruihua Song, Xin Gao, Tao Xiang, et al. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, 13(1):3094, 2022.

Xidong Feng, Yicheng Luo, Ziyan Wang, Hongrui Tang, Mengyue Yang, Kun Shao, David Mguni, Yali Du, and Jun Wang. Chessgpt: Bridging policy learning and language modeling. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 7216–7262. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/16b14e3f288f076e0ca73bdad6405f77-Paper-Datasets_and_Benchmarks.pdf.

Grant Fergusson, Caitriona Fitzgerald, Chris Frascella, Megan Iorio, Tom McBrien, Calli Schroeder, Ben Winters, and Enid Zhou. Generating harms: Generative ai's impact & paths forward. *Electronic Privacy Information Center*, 2023.

Yingqiang Ge, Wenyue Hua, Kai Mei, jianchao ji, Juntao Tan, Shuyuan Xu, Zelong Li, and Yongfeng Zhang. Openagi: When llm meets domain experts. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 5539–5568. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1190733f217404edc8a7f4e15a57f301-Paper-Datasets_and_Benchmarks.pdf.

R Gibson and Margaret Barnes. Evolution and ecology of cleaning symbioses in the sea. *Oceanography and marine biology: an annual review*, 38:311, 2000.

Charles AE Goodhart. *Problems of monetary management: the UK experience*. Springer, 1984.

Sagar Goyal, Eti Rastogi, Sree Prasanna Rajagopal, Dong Yuan, Fen Zhao, Jai Chintagunta, Gautam Naik, and Jeff Ward. Healai: A healthcare llm for effective medical documentation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, pp. 1167–1168, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703713. doi: 10.1145/3616855.3635739. URL https://doi.org/10.1145/3616855.3635739.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2020.

Matthew Ho, Aditya Sharma, Justin Chang, Michael Saxon, Sharon Levy, Yujie Lu, and William Yang Wang. Wikiwhy: Answering and explaining cause-and-effect questions. In *The Eleventh International Conference on Learning Representations*, 2023.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Todd Houghton, Michael Saxon, Zeming Song, Hoa Nyugen, Hanqing Jiang, and Hongbin Yu. 2d grating pitch mapping of a through silicon via (tsv) and solder ball interconnect region using laser diffraction: Ieee electronic components and technology conference, 2016. In *2016 IEEE 66th Electronic Components and Technology Conference (ECTC)*, pp. 2222–2227. IEEE, 2016.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20406–20417, 2023.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024.

Jaap Jumelet and Dieuwke Hupkes. Do language models understand anything? on the ability of LSTMs to understand negative polarity items. In Tal Linzen, Grzegorz Chrupała, and Afra Alishahi (eds.), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 222–231, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5424. URL https://aclanthology.org/W18-5424.

Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Kaya Stechly, Mudit Verma, Siddhant Bhambri, Lucas Saldyt, and Anil Murthy. Llms can't plan, but can help planning in llm-modulo frameworks. *arXiv preprint arXiv:2402.01817*, 2024.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15696–15707. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/kandpal23a.html.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4110–4124, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.324. URL https://aclanthology.org/2021.naacl-main.324.

Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pp. 193–203, 2023.

Ann Elizabeth Fowler La Berge. The history of science and the history of microscopy. *Perspectives on Science*, 7(1):111–142, 1999.

Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, et al. Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, 34:29348–29363, 2021.

David Leverington. *A History of Astronomy: from 1890 to the Present*. Springer Science & Business Media, 2012.

Lei Li, Jingjing Xu, Qingxiu Dong, Ce Zheng, Xu Sun, Lingpeng Kong, and Qi Liu. Can language models understand physical concepts? In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11843–11861, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.726. URL https://aclanthology.org/2023.emnlp-main.726.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=iO4LZibEqW. Featured Certification, Expert Certification.

Q Vera Liao and Ziang Xiao. Rethinking model evaluation as narrowing the socio-technical gap. *arXiv preprint arXiv:2306.03100*, 2023.

Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. Are we learning yet? a meta review of evaluation failures across machine learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Mark Liberman. Obituary: Fred jelinek. *Computational Linguistics*, 36(4):595–599, 2010.

Bill Yuchen Lin, Khyathi Chandu, Faeze Brahman, Yuntian Deng, Abhilasha Ravichander, Valentina Pyatkin, Ronan Le Bras, and Yejin Choi. Wildbench: Benchmarking llms with challenging tasks from real users in the wild, 2024. URL https://huggingface.co/spaces/allenai/WildBench.

Michael Xieyang Liu, Advait Sarkar, Carina Negreanu, Benjamin Zorn, Jack Williams, Neil Toronto, and Andrew D. Gordon. "what it wants me to say": Bridging the abstraction gap between end-user programmers and code-generating large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3580817. URL https://doi.org/10.1145/3544548.3580817.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KUNzEQMWU7.

Christopher D Manning. Human language understanding & reasoning. *Daedalus*, 151(2): 127–138, 2022.

EM Mansoor. The application of probability to tolerances used in engineering designs. *Proceedings of the Institution of Mechanical Engineers*, 178(1):29–39, 1963.

Joshua Maynez, Priyanka Agrawal, and Sebastian Gehrmann. Benchmarking large language model capabilities for conditional generation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9194–9213, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.511. URL https://aclanthology.org/2023.acl-long.511.

Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334.

Kyle Melnick. Air canada chatbot promised a discount. now the airline has to pay it. *The Washington Post*, 2024. URL https://www.washingtonpost.com/travel/2024/02/18/air-canada-airline-chatbot-ruling/.

Julian Michael, Ari Holtzman, Alicia Parrish, Aaron Mueller, Alex Wang, Angelica Chen, Divyam Madaan, Nikita Nangia, Richard Yuanzhe Pang, Jason Phang, et al. What do nlp researchers believe? results of the nlp community metasurvey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16334–16368, 2023.

Melanie Mitchell and David C Krakauer. The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120, 2023.

Atsushi Mizumoto and Masaki Eguchi. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050, 2023. ISSN 2772-7661. doi: https://doi.org/10.1016/j.rmal.2023.100050. URL https://www.sciencedirect.com/science/article/pii/S2772766123000101.

Jared Moore. Language models understand us, poorly. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 214–222, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.16. URL https://aclanthology.org/2022.findings-emnlp.16.

Meredith Ringel Morris, Jascha Sohl-dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. Levels of agi: Operationalizing progress on the path to agi. *arXiv preprint arXiv:2311.02462*, 2023.

Hugo Neri and Fabio Cozman. The role of experts in the public perception of risk of artificial intelligence. *AI & society*, 35:663–673, 2020.

NIST. AI Risk Management Framework: Second Draft. Technical report, National Institute for Standards and Technology, 09 2022. URL https://www.nist.gov/system/files/documents/2022/08/18/AI_RMF_2nd_draft.pdf.

NIST. ARTIFICIAL INTELLIGENCE SAFETY INSTITUTE CONSORTIUM COOPERATIVE RESEARCH AND DEVELOPMENT AGREEMENT. Technical report, National Institue for Standards and Technology, 12 2023. URL https://www.nist.gov/system/files/documents/2023/12/15/AISIC%20FINAL%20APPROVED%20TEMPLATE_FINAL%20FINAL%2012152023%20reference%20copy.pdf.

Juhyun Oh, Eunsu Kim, Inha Cha, and Alice Oh. The generative ai paradox on evaluation: What it can solve, it may not evaluate. *arXiv preprint arXiv:2402.06204*, 2024.

Scott W O'Leary-Kelly and Robert J Vokurka. The empirical assessment of construct validity. *Journal of operations management*, 16(4):387–405, 1998.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Garson O'Toole. *Hemingway didn't say that: The truth behind familiar quotations*. Brilliance Audio, 2017.

Simon Ott, Adriano Barbosa-Silva, Kathrin Blagec, Jan Brauner, and Matthias Samwald. Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nature Communications*, 13(1):6793, 2022.

Sungjin Park, Seungwoo Ryu, and Edward Choi. Do language models understand measurements? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1782–1792, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.128. URL https://aclanthology.org/2022.findings-emnlp.128.

Keiran Paster. Testing language models on a held-out high school national finals exam. https://huggingface.co/datasets/keirp/hungarian_national_hs_finals_exam, 2023.

Zhencan Peng, Zhizhi Wang, and Dong Deng. Near-duplicate sequence search at scale for large language model memorization evaluation. *Proc. ACM Manag. Data*, 1(2), jun 2023. doi: 10.1145/3589324. URL https://doi.org/10.1145/3589324.

P Jonathon Phillips, Hyeonjoon Moon, Syed A Rizvi, and Patrick J Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on pattern analysis and machine intelligence*, 22(10):1090–1104, 2000.

Steven Piantadosi and Felix Hill. Meaning without reference in large language models. In *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)*, 2022. URL https://openreview.net/forum?id=nRkJEwmZnM.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pp. 180–191, 2018.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*, 2024.

Christopher Potts. Is it possible for language models to achieve language understanding? Medium post, 2020. URL https://chrisgpotts.medium.com/is-it-possible-for-language-models-to-achieve-language-understanding-81df45082ee2.

Inioluwa Deborah Raji, Emily Denton, Emily M Bender, Alex Hanna, and Amandalynne Paullada. Ai and the everything in the whole wide world benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 464–483. IEEE, 2023.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Abdullahi Saka, Ridwan Taiwo, Nurudeen Saka, Babatunde Abiodun Salami, Saheed Ajayi, Kabiru Akande, and Hadi Kazemi. Gpt models in construction industry: Opportunities, limitations, and a use case validation. *Developments in the Built Environment*, 17:100300, 2024. ISSN 2666-1659. doi: https://doi.org/10.1016/j.dibe.2023.100300. URL https://www.sciencedirect.com/science/article/pii/S2666165923001825.

Lucie H Salwiczek, Laurent Prétôt, Lanila Demarta, Darby Proctor, Jennifer Essler, Ana I Pinto, Sharon Wismer, Tara Stoinski, Sarah F Brosnan, and Redouan Bshary. Adult cleaner wrasse outperform capuchin monkeys, chimpanzees and orang-utans in a complex foraging task derived from cleaner–client reef fish cooperation. *PLoS One*, 7(11):e49068, 2012.

Michael Saxon, Xinyi Wang, Wenda Xu, and William Yang Wang. PECO: Examining single sentence label leakage in natural language inference datasets through progressive evaluation of cluster outliers. In Andreas Vlachos and Isabelle Augenstein (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3061–3074, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.223. URL https://aclanthology.org/2023.eacl-main.223.

Charles Singer. Notes on the early history of microscopy. *Proceedings of the Royal Society of Medicine*, 7(Sect_Hist_Med):247–279, 1914.

Prasann Singhal, Jarad Forristal, Xi Ye, and Greg Durrett. Assessing out-of-domain language model performance from few examples. In *Conference of the European Chapter of the Association for Computational Linguistics*, 2022. URL https://api.semanticscholar.org/CorpusID:252872900.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swkedrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel Stern Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone

Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, vinay uday prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=uyTL5Bvosj.

Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. On the self-verification limitations of large language models on reasoning and planning tasks. *arXiv preprint arXiv:2402.08115*, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Ken Thompson. Reflections on trusting trust. *Communications of the ACM*, 27(8):761–763, 1984.

Songül Tolan, Annarosa Pesole, Fernando Martínez-Plumed, Enrique Fernández-Macías, José Hernández-Orallo, and Emilia Gómez. Measuring the occupational impact of ai: tasks, cognitive abilities and ai benchmarks. *Journal of Artificial Intelligence Research*, 71: 191–236, 2021.

Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. Large language models still can't plan (a benchmark for LLMs on planning and reasoning about change). In *NeurIPS 2022 Foundation Models for Decision Making Workshop*, 2022. URL https://openreview.net/forum?id=wUU-7XTL5XO.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*, 2023a.

Xinyi Wang, Wenhu Chen, Michael Saxon, and William Yang Wang. Counterfactual maximum likelihood estimation for training deep networks. *Advances in Neural Information Processing Systems*, 34:25072–25085, 2021.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, et al. Factcheck-gpt: End-to-end fine-grained document-level fact-checking and correction of llm output. *arXiv preprint arXiv:2311.09000*, 2023b.

Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, et al. The generative ai paradox:"what it can create, it may not understand". In *The Twelfth International Conference on Learning Representations*, 2023.

Xixi Xu, Zhongang Qi, Jianqi Ma, Honglun Zhang, Ying Shan, and Xiaohu Qie. Bts: A bi-lingual benchmark for text segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19152–19162, June 2022.

Linyi Yang, Shuibai Zhang, Libo Qin, Yafu Li, Yidong Wang, Hanmeng Liu, Jindong Wang, Xingxu Xie, and Yue Zhang. Glue-x: Evaluating natural language understanding models from an out-of-distribution generalization perspective. *ArXiv*, abs/2211.08073, 2022. URL https://api.semanticscholar.org/CorpusID:253523094.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 535–546, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.45. URL https://aclanthology.org/2021.naacl-main.45.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 5484–5505. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/117c5c8622b0d539f74f6d1fb082a2e9-Paper-Datasets_and_Benchmarks.pdf.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. (inthe) wildchat: 570k chatgpt interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*, 2023a.

Zirui Zhao, Wee Sun Lee, and David Hsu. Large language models as commonsense knowledge for large-scale task planning. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 31967–31987. Curran Associates, Inc., 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/65a39213d7d0e1eb5d192aa77e77eeb7-Paper-Conference.pdf.

Aojun Zhou, Ke Wang, Zimu Lu, Weikang Shi, Sichun Luo, Zipeng Qin, Shaoqing Lu, Anya Jia, Linqi Song, Mingjie Zhan, and Hongsheng Li. Solving challenging math word problems using GPT-4 code interpreter with code-based self-verification. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=c8McWs4Av0.

Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, and Maarten Sap. Is this the real life? is this just fantasy? the misleading success of simulating social interactions with llms, 2024b.